Thesis proposal: COGNILENS: Analyzing Cognitive Decline in Language Models for Alzheimer's Monitoring

Jonathan Guerne[1,2],
[1] University of Neuchâtel, Switzerland
[2]Haute Ecole Arc Ingénierie,
University of Applied Sciences and Arts Western Switzerland (HES-SO)

Abstract

This research proposal describes a cross-disciplinary project aimed at developing Digital Twins (DTs) of Alzheimer's Disease (AD) using Language Models (LMs). By mimicking the functional deficits observed in individuals with AD, these DTs will serve as tools for early detection and understanding of disease progression. Several approaches to altering the LM will be explored, and the resulting effects on brain score — an evaluation of the correlation between brain activity and the LM's internal activations — will be studied. Detection models will be trained based on each approach; these models will be compared against themselves and the state-of-the-art. Two converging lines of evidence motivate this work: LMs achieve high accuracy in classifying AD from speech transcripts, and their internal representations correlate significantly with human brain activity during language processing. If successful, this project could lead to significant advancements in the early detection and monitoring of AD, ultimately improving patient outcomes.

1 Introduction

Alzheimer's Disease (AD) is a neurodegenerative disorder primarily affecting the elderly, characterized by memory loss and cognitive decline (Mandell and Green, 2011). Despite decades of research, AD and, by extension, Mild Cognitive Impairment (MCI) remain difficult to detect before late stages, hindering treatment efficacy. Developing detection methods and monitoring tools is key to improving patient outcomes (Frisoni et al., 2021). Numerous solutions spanning different fields have already been explored. Neurobiologically informed approaches focus on the analysis of brain activity to distinguish notable effects of the disease (Alarjani and Almarri, 2024). Natural Language Processing (NLP) approaches focus on the study of language impairment, which has been shown to be relevant even at an early stage (Verma and Howard, 2012). In addition, screening techniques meant to rapidly obtain an assessment of cognitive function have been developed, such as Mini-Mental State Examination (MMSE) (Arevalo-Rodriguez et al., 2021) or Montreal Cognitive Assessment (MoCA) (Nasreddine et al., 2005).

NLP approaches using transformer techniques reach accuracy as high as 85% on transcript (purely text-based) classification between healthy controls and individuals with AD. Huth et al. (2016); Schrimpf et al. (2020); Caucheteux (2023) recently established that for a given language task, activations inside Artificial Neural Networks (ANNs) of Transformers resemble the neural activity in the human brain. However, it remains unclear whether the internal representations of Transformer-based detection models bear any meaningful resemblance to the neurocognitive alterations observed in individuals with AD. This gap is significant: if LMs trained on AD data encode changes that are aligned with human neural activity, they could serve not only as classifiers but as interpretable models of cognitive degradation, potentially contributing to our understanding of the disease. On the other hand, if their success is purely statistical, relying on textual patterns without deeper cognitive alignment, it opens new opportunities to refine these models for better interpretability and clinical relevance.

Yet, at present, there is no empirical evidence connecting the internal state changes of high-performing AD detection models to the known functional alterations of the human brain activity seen in AD (Greicius et al., 2004; Verma and Howard, 2012). This disconnect raises a fundamental question: do current LMs that succeed in AD detection actually simulate the disease in any cognitively meaningful way?

This study proposes a novel method for early detection and progression monitoring of AD by using altered LMs. We hypothesize that if we alter LMs such that their activation patterns mimic

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 4: Student Research Workshop, pages 170–181

the neural activity patterns of individuals with AD, they will also share similar language deficits and vice versa. From this perspective, our goal is therefore to create Digital Twins (DTs) that accurately model the activation function of the individual's brain and could be used to detect, monitor, and open opportunity windows to treat AD. The LM's cognitive degradation will be monitored via the MoCA screening test. Brain score (Schrimpf et al., 2021) will be used to monitor the similarity between brain and ANN.

We expect to achieve competitive performance in the detection of AD while offering new monitoring capabilities. We aim to match and potentially outperform the current state-of-the-art. Partially altered models (see Figure 3) will be leveraged to improve the detection pipeline.

1.1 Research Questions

- RQ1: To what extent can a LM's cognitive and linguistic performance be evaluated to characterize model degradation during alteration?
- RQ2: To what extent does simulating cognitive decline through network alterations affect LM performance, and how does this relate to AD detection capability?
- RQ3: Can altered LM help monitor the progression of AD for specific individuals?

2 Related Work

2.1 AD and MCI screening

To provide quantifiable targets for automatically detecting AD and MCI, researchers notably rely on cognitive screening techniques, such as MMSE or MoCA. MMSE is designed to assess the overall mental function of the patient (Arevalo-Rodriguez et al., 2021); it is used to screen for cognitive impairment and track changes over time. It consists of a questionnaire evaluating key cognitive domains, including orientation, attention, memory, language, and visual-spatial skills. It is a widely used approach yet it has known limitations: it cannot be trusted to detect MCI; in other words, it cannot be trusted for early AD detection. MoCA was proposed as a response to this limitation as it is better suited to screen patients with mild cognitive complaints (Nasreddine et al., 2005). It follows the same format as MMSE with a total of 30 questions and as many available points. Nevertheless, neither of these tests is sufficient to establish a diagnosis of AD or MCI on their own; they are meant to be used as part of a comprehensive assessment that includes clinical evaluation, medical history, and other diagnostic tests.

2.2 Language-based detection of AD

DementiaBank[1] is one of the most widely used collections of datasets for AD detection from linguistic data. One of its most notable entries, the ADReSS (Luz et al., 2020) challenge, provides a standardized benchmark for evaluating detection models. The ADReSS dataset contains 156 audio recordings of picture descriptions from both individuals with AD and healthy controls, along with their corresponding transcripts, demographic information, and cognitive scores (MMSE). The challenge has attracted significant attention from the research community, leading to the development of various models; it is now widely used across studies as a shared benchmark (Luz et al., 2021).

Various studies have explored the use of NLP techniques to analyze speech and text data from individuals with AD (Qi et al., 2023; Yang et al., 2022). They have shown that certain linguistic features, such as semantic impairment, acoustic abnormality, syntactic impairment, and information impairment, can be used to distinguish between individuals with AD and healthy controls (Fraser et al., 2016; Thomas et al., 2005). Researchers have also highlighted the potential of using verbal utterances to detect MCI (Padhee et al., 2020; König et al., 2015; Hernández-Domínguez et al., 2018).

Deep Learning (DL) models are becoming the new standard for most NLP tasks. More specifically, the Transformer architecture proposed by Vaswani et al. (2023) introduced new capabilities for processing sequential data that led to significant improvements in virtually all NLP benchmarks. These models are based on the attention mechanism, which allows them to capture the context and meaning of words in a sentence more effectively compared to previous recurrent architectures such as LSTM (Hochreiter and Schmidhuber, 1997) or GRU (Cho et al., 2014). The Transformer architecture has been used to develop a variety of LMs such as BERT (Devlin et al., 2019) or GPT-2 (Radford et al., 2019) and is still the basis of most recent models.

To avoid the need to train LMs from scratch, researchers relied on fine-tuning, a strategy that

[1] https://dementia.talkbank.org/

limits the scale of the required training set by relying on a pretrained model. Fine-tuning has been leveraged to specialize existing models in learning features relevant to speech in individuals with AD (Balagopalan et al., 2020, 2021; Pan et al., 2021; Ding et al., 2024; Chi et al., 2025). Balagopalan et al. (2020) demonstrated that a fine-tuned BERT model outperformed traditional Machine Learning (ML) models on the ADReSS test set, achieving accuracies of 83.3% and 81.3%, respectively. Li et al. (2022a) proposed GPT-D, a GPT-2-based model that improved the state-of-the-art accuracy to 85% — our target to match.

2.3 Neuroimaging-based detection of AD

Researchers aim to identify biomarkers that can aid in the diagnosis and monitoring of disease progression. functional Magnetic Resonance Imaging (fMRI) is an imaging technique that measures brain activity by detecting changes in blood flow. It is based on the principle that when a brain region is more active, it consumes more oxygen, leading to an increase in blood flow to that region. This change can be detected, allowing the mapping of brain activity in response to various tasks or stimuli (Logothetis et al., 2001). Compared to other neuroimaging techniques, fMRI is (I) functional, implying that it can be used to monitor changes in activity in the brain, not its structure, and (II) time-based, meaning that it captures 4-dimensional data (3D space + time), allowing the capture of the dynamic aspect of brain activity. In the context of AD, functional changes are expected to appear before structural changes, making fMRI more relevant for early detection of the disease (Dennis and Thompson, 2014).

The use of Magnetic Resonance Imaging (MRI) (including fMRI) to study the brain activity of individuals with AD has been explored for a couple of decades (Grossman et al., 2003; Domoto-Reilly et al., 2012; Alarjani and Almarri, 2024). Some regions of the brain are of particular interest when studying AD. As an example, the Default Mode Network (DMN) (Greicius et al., 2004) is a network of brain regions that are active when the brain is at rest and not focused on the outside world. It includes the medial prefrontal cortex, posterior cingulate cortex, and angular gyrus. Its disruption has been linked to cognitive decline in AD. The anterior temporal lobe (ATL) (Verma and Howard, 2012) is involved in semantic memory and language processing; its dysfunction is associated with language impairments observed in individuals with AD.

Cha et al. (2013) studied the functional alteration patterns of the DMN in normal aging, amnestic Mild Cognitive Impairment (aMCI), and AD. They found that the DMN showed significant functional alterations in both aMCI and individuals with AD compared to individuals with normal aging, yet these alterations were more pronounced, and sometimes unique to, individuals with AD compared to individuals with aMCI. This implies that the functional alterations in the DMN could serve as potential biomarkers for distinguishing between normal aging, aMCI, and AD.

2.4 LMs as Cognitive Models

Huth et al. (2016); Schrimpf et al. (2020); Caucheteux (2023) investigated the use of LMs as cognitive models, suggesting that transformer-based architectures can capture aspects of human cognition. Schrimpf et al. (2021) demonstrated that leading transformer models can account for nearly all explainable variance in neural responses to sentences, generalizing across multiple datasets and neuroimaging modalities such as fMRI and electroencephalography (EEG). In other words, the patterns of activity observed in human brains can be almost perfectly predicted by the activations within transformer-based models. This means that the model's internal representations of words, syntax, and meaning align with the neural representations observed in human language areas. The fact that this generalizes across different datasets and measurement techniques indicates that these models capture, to some degree, biologically relevant principles of language. Nevertheless, it is important to note that while these models can predict neural responses, they do not replicate the full complexity of human brain function. Thus, there are some limitations to their use as cognitive models that still need to be explored and documented (Gauthier and Levy, 2019; Caucheteux and King, 2021).

To measure the similarity between LM and brain activity, researchers introduced the concept of "brain score", which quantifies the correlation between the activations of an ANN and the neural activity recorded from the brain when processing the same sentences (Schrimpf et al., 2018). It is computed by fitting a linear model to predict the brain activity of one Region of Interest (ROI) — defined as a set of voxels which are the smallest units of analysis in MRI — given the activations of an ANN as input (see Figure 1). This

approach allows researchers to identify which layers of the LM are most similar to brain activity and to explore how different architectural choices affect this similarity. Brain score tests how well a simple linear mapping from these layer representations can predict neural activity recorded during the same language input. Empirically, predictivity tends to improve from early to middle layers and then plateaus or declines (Schrimpf et al., 2021; Caucheteux et al., 2022). Moreover, different layers emphasize distinct linguistic functions: syntax-dominant information aligns more with superior temporal regions, whereas semantic/compositional information aligns with inferior-frontal and parietal areas and extends toward the ATL. To obtain such results, Caucheteux et al. (2021) had to demonstrate that the semantic and syntactic components of GPT-2 activations could be isolated and that their brain scores could be computed separately.

2.5 Digital Twins in Biomedical Research

Laubenbacher et al. (2024) provides a comprehensive overview of the concept of DTs in biomedical research. They define a DT as a virtual representation of a physical entity that can be used for simulation, analysis, and optimization. DTs can be used to model complex biological systems, such as organs or diseases, and to simulate their behavior under different conditions. DTs have the potential to improve our understanding of biological systems and to develop new treatments and therapies. Wu and Koelzer (2024); Ashraf et al. (2024) explore the use of generative models, such as GPT-2, as DTs in biomedical research, yet the specific use of LMs as brain activity models is still to be explored, especially in the context of AD.

GPT-2 has been shown to yield its highest brain predictivity within the bilateral superior and middle temporal cortices, extending into the ATL (Caucheteux et al., 2021, 2022). Thus, this region is simultaneously a clinically relevant target for monitoring emerging semantic impairment and a locus where modern predictive language models achieve strong brain-model correspondence, highlighting the potential of LMs as DTs of brain activity in the context of AD.

LM could also be used to pass screening tests designed for human patients, as demonstrated by Dayan et al. (2024) when they highlighted the potential of such tests to showcase the relative cognitive performance of different LMs.

3 Proposed Methodology

3.1 Research Design

The research will follow a structured framework repeated for each alteration experiment (see Figure 2). An alteration approach will be implemented with the goal of altering the LM in ways that lead to cognitive decline. The cognitive decline will be monitored with the assessment module by quantifying the model's cognitive performance at different stages of the alteration. The analysis will focus more on the relative degradation of cognitive performance rather than the absolute performance. This allows us to compare different models while limiting the impact of the model's initial performance or lack thereof. It is crucial that this alteration is gradual; otherwise the cognitive decline would be too great and would render the study meaningless.

A detection pipeline will be implemented to leverage the altered models. This pipeline will serve to compare the different alteration approaches against each other and the state-of-the-art. Each model will be saved at different stages, or checkpoints, of their alteration (see Figure 3).

We will conclude the study with a preliminary study of a DT for AD using a longitudinal dataset. The goal will be to monitor the progression of AD for specific individuals.

3.2 Data and Tools

Data The *Narratives* dataset (Nastase et al., 2021), containing fMRI recordings of 345 subjects listening to 27 stories, will be used primarily for brain-score mapping and normative modeling from healthy subjects. This will be supplemented by the *Petit Prince fMRI collections* (Li et al., 2022b; Momenian et al., 2024), providing story-listening fMRI from young and elderly healthy adults for aging-related analyses and neurobiologically informed alteration. For transcript-based alteration and evaluation, we will use the *ADReSS* dataset (Luz et al., 2020), which contains picture-description transcripts with a benchmark split. The *Delaware corpus* (Lanzi et al., 2023) provides multi-task discourse transcripts, audio, and cognitive scores to augment *ADReSS*. For longitudinal studies, we will use *ADReSSo* (Luz et al., 2021) and *Pitt Corpus* (Becker et al., 1994), which contain speech/transcript datasets with MMSE as cognitive assessments. The *Baycrest corpus* (Kielar et al., 2016) provides narratives from individuals with MCI/AD with MoCA and resting-state fMRI.

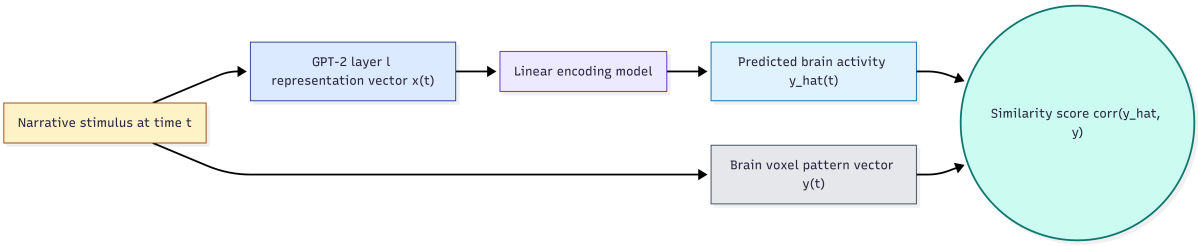

Figure 1: Illustration inspired from (Caucheteux et al., 2022), measures the mapping between the subject's brain activations and the activations of GPT-2, both elicited by the same narrative. To this end, a linear model is fitted to predict the brain activity of one voxel Y, given GPT-2 activations X as input. The degree of mapping is called "brain score". Brain scores can be averaged across fMRI voxels, and different layers of GPT-2.

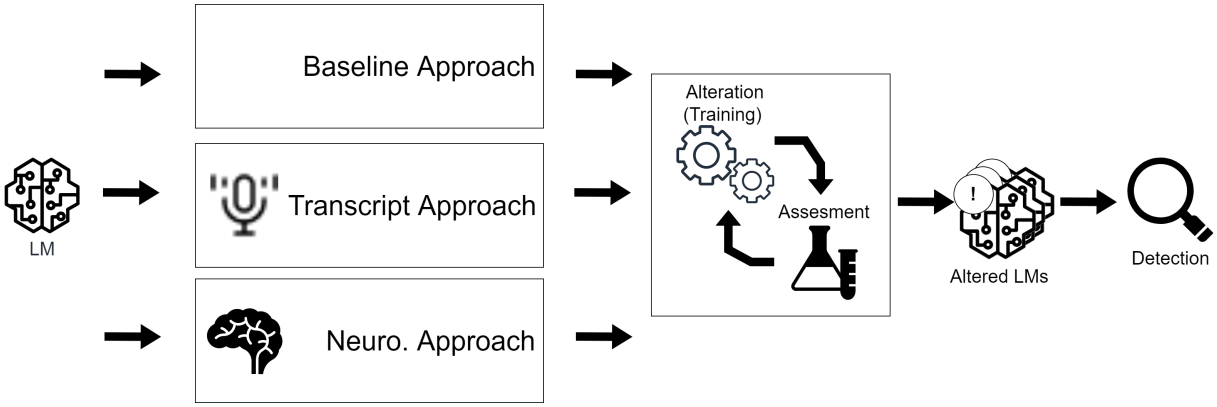

Figure 2: The research is designed as an iterated framework. An alteration approach is implemented with the goal of altering the LM in ways that lead to cognitive decline. The effect of the alteration is monitored with the assessment module and brain score computation. A detection pipeline is implemented to leverage the altered models.

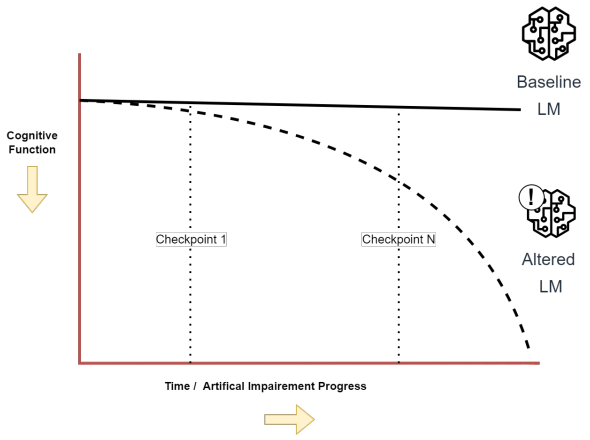

Figure 3: An altered LM is expected to lose its cognitive performance over the alteration process. This degradation is expected to be gradual, allowing us to monitor the evolution of brain score and cognitive performance at different stages or "checkpoints".

Finally, the *VAS corpora* (Liang et al., 2022) and *Connected-speech benchmark* (Luz et al., 2024) provide speech datasets with MoCA/MMSE for optional acoustic/hybrid detection experiments.

Tools The primary tools for our research include *MoCA* (Nasreddine et al., 2005) as the basis for the assessment module; *brain score/neural alignment* tools implementing the linear-mapping brain score following Schrimpf et al. (2021) (considering existing brain-score toolkits where compatible[2]); and *Hugging Face*[3] for access to pre-trained LMs and the Transformers library for model loading and fine-tuning.

3.3 Implementation Plan

3.3.1 Cognitive and Linguistic Assessment Module

We will develop a dual-perspective assessment module to quantify cognitive and linguistic changes in LMs during alteration, enhancing measurement

[2] https://brain-score-language.readthedocs.io/en/latest/

[3] https://huggingface.co/

sensitivity and providing redundancy if one perspective fails to capture meaningful changes.

Cognitive Performance Assessment We will adapt the MoCA to evaluate model responses across cognitive domains. This approach builds on Dayan et al. (2024), who compared Large Language Models (LLMs) using MoCA, and Binz and Schulz, who showed LLMs can solve cognitive tasks comparably to humans. The remote administration paradigm (Wong et al., 2015) simplifies adaptation for text-based LMs.

Some questions will require adaptation or exclusion due to modality restrictions, with reference models (e.g., ChatGPT) used to validate task feasibility before inclusion. Model selection will balance practical advantages of smaller models and their extensive brain alignment literature (Caucheteux et al., 2022) against the need for adequate baseline performance[4].

Linguistic Performance Assessment We will monitor three complementary linguistic metrics to capture different facets of language degradation. **Perplexity** (Goodman, 2001) quantifies predictive uncertainty, with increases signaling degraded linguistic capability. **N-gram diversity** (Li et al., 2016) measures lexical richness, relevant as reduced diversity characterizes AD language production (Williams et al., 2021). **Bert-score** (Zhang et al., 2020) captures semantic preservation using contextual embeddings, addressing the semantic impairments central to AD.

Integrated Framework This dual assessment investigates whether cognitive and linguistic deterioration follow parallel trajectories during alteration — key to determining if altered models genuinely simulate AD-like decline. Cognitive scores provide interpretable, domain-specific measures aligned with clinical paradigms, while linguistic metrics offer continuous indicators that may reveal subtle changes not captured by discrete tasks. Importantly, this dual-perspective approach provides methodological redundancy: if one assessment dimension proves insufficiently sensitive to capture model degradation, the other may still provide meaningful indicators. Together, they strengthen the robustness of our evaluation framework.

[4]The brain score language leaderboard could be a useful resource for this comparison (see https://www.brain-score.org/language/leaderboard/)

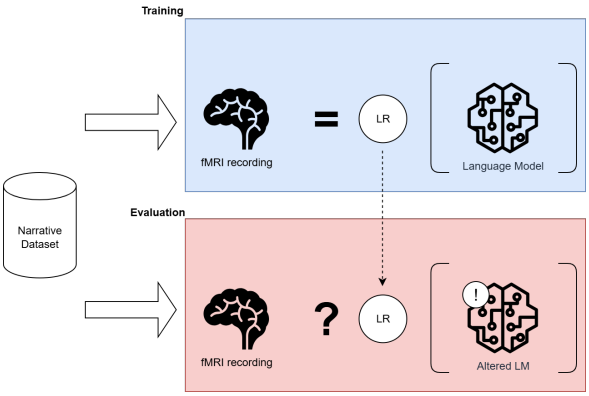

Figure 4: We use the brain score framework described in Section 2.4. The trained linear model is applied to the altered model to compute the brain score delta (the difference between the brain score of the altered model and the original model).

3.3.2 Brain score computation and normative modeling

Our goal is to study the LM activation and explore the correlation between the LM's loss of cognitive performance and the deviation from healthy brain activity. To do this, we will rely on Schrimpf et al. (2021) to compute the linear model required for brain score computation. We will follow the training strategy proposed by Schrimpf et al. (2021) which involves a hold-out validation set of 20% of the dataset and a normalization of the predictivity scores per ROI. Inspired by the work of Rutherford et al. (2022), we will implement a normative model. The idea of normative modeling is to monitor the behaviors of a single entity or observation against the expected norm represented by the normative model. In the case of this proposal, the model will be used to monitor the change in brain activity after the alteration process. We will use the "Narratives" dataset (Nastase et al., 2021) to fit the normative model. Once trained, the model will be used to compute the brain score of the altered model, enabling us to compute the delta of brain score (see Figure 4).

3.3.3 Model alteration approaches

We designed an empirical framework where different alteration approaches will be compared against each other: a baseline approach based on neuron dropout, an approach leveraging transcripts from individuals with AD to fine-tune LMs, and a neurobiologically inspired approach. All alteration approaches will share a similar implementation structure; crucially, they will all output *multiple*

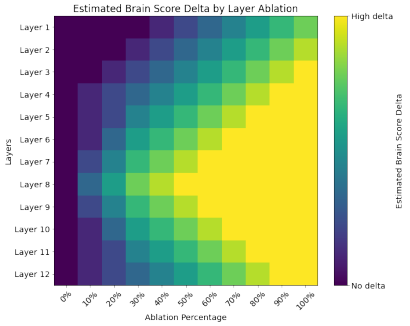

Figure 5: We expect the brain score to degrade as the model is altered; this degradation might vary depending on the layer. This illustration assumes a GPT-2 model with 12 layers will be used.

versions of the altered model (see Figure 3). The multiplication of outputs allows us to study the nuances in the evolution of cognitive performance and brain score, ultimately leading to a better detection pipeline. The comparative nature of this framework ensures that even if individual approaches yield weaker-than-expected effects, the differential patterns across methods will provide insights into which types of changes in LMs are more or less effective at simulating cognitive decline, and which evaluation methods are most sensitive to these changes. Overall, the alteration process will share strong similarities with the common training stage of ML, the difference being that the alteration process aims to degrade the model's performance (in a specific way) rather than improve it. There is a risk that the capabilities of the selected LM will drop so significantly that it will render all tasks impossible to complete. Our aim will be to limit this risk with small, gradual, alterations.

Baseline Alteration Approach This alteration approach will serve as a baseline for the approaches described in the following sections.

The inner connections of neurons within the LM will be randomly and gradually removed. This is inspired by a common training regularization technique called dropout (Srivastava et al., 2014). We hypothesize that the introduced lesion could lead to a degradation of the model's linguistic capabilities similar to those observed in individuals with AD. Dropout could also be seen as a biomimetic simulation of known synaptic degradation in certain areas of the brain (Terry et al., 1991). The viability of this approach is also strengthened by the work of Li et al. (2022a), who followed a similar methodology to create GPT-D, an altered version of GPT-2 used to reach state-of-the-art performance in AD classification from textual input. We will monitor the brain score of the model at different stages of the alteration process; we expect to observe a brain score delta that varies depending on the layer being altered and the degree of ablation (see Figure 5).

Transcript-Based Alteration Approach This approach fine-tunes the LM on ADReSS transcripts (Luz et al., 2020) to shift its output toward language patterns associated with AD (lexical restriction, simplified structure, repetition). Unlike common AD detection pipelines that attach or fine-tune a classifier over the LM (Balagopalan et al., 2020, 2021; Pan et al., 2021; Ding et al., 2024), we keep the original next-token objective. We will use mixed batches of healthy and AD transcripts with a controlled increase in the AD weighting over time. This gradual fine-tuning aims to slowly shift the model's linguistic capabilities toward those observed in AD while preserving some of its original functionality. Fine-tuning will stop once significant cognitive degradation is observed via the assessment module.

This transcript-driven adaptation improves on the random ablation baseline by providing a more structured and data-informed approach to model alteration, potentially leading to more realistic and clinically relevant cognitive decline patterns. It also offers a complementary perspective to the neurobiologically inspired approach, which directly targets model internals based on brain activity patterns rather than behavioral output.

Neurobiologically Inspired Alteration Approach We propose a neurobiologically informed alteration of LMs that is rooted in disease-related brain dysfunction patterns rather than behavioral output. AD-specific effects and ROIs in the brain will be derived from the literature and (if available) fMRI. As a first assessment, the following could serve as a starting ground for our research: ATL semantic hub degeneration (Grossman et al., 2003; Domoto-Reilly et al., 2012; Ralph et al., 2017); disrupted DMN connectivity (Cha et al., 2013); synapse loss and dysfunction (Terry et al., 1991; Pelucchi et al., 2022). Once we have identified the relevant ROIs, we will use brain score to compute layer similarity between LM internal states and region-specific activity. The designated layer could then be targeted for alteration; for instance, synaptic lesion could be simulated.

This approach will share similarities with the

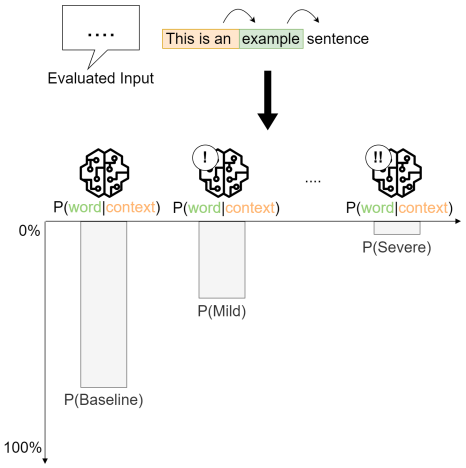

Figure 6: To detect AD, a text input will be used to compute the perplexity of both the baseline and altered model. This approach provides a certainty rating.

baseline approach, but the alteration will be more targeted and informed by neurobiological data. The goal is to create a model that not only exhibits cognitive decline but does so in a manner that reflects known patterns of brain dysfunction in AD.

3.3.4 Detection pipeline

Developed models will be compared in a classification task with the goal of properly identifying the presence (or absence) of the disease from textual input.

Perplexity-based detection Perplexity is an evaluation metric used with LMs (Goodman, 2001). It computes the likelihood that a LM would have to generate a certain series of words. The smaller the perplexity, the more likely it is that the model would have produced this exact sentence, and vice versa. In this scenario, we will compute the perplexity of multiple models, each representing a specific point in the alteration process. The benefit of using multiple models at once is that we will be able to determine for which model the perplexity is smallest and, by extension, which step in the alteration process it represents. We will aggregate the perplexity per token of the input text for each model. To determine the most likely stage of the disease, we will convert these values into a probability distribution using a softmax function (see Figure 6). This approach enables us to address whether effective detection requires different model characteristics than those that emerge from simulating cognitive decline; in that sense, null or weak relationships between alteration degree and detection performance would still provide valuable insights into the limitations of the proposed framework and could guide future research toward other alteration or detection strategies.

Inner network activation-based detection While perplexity measures output-level surprise, this approach examines internal processing stability by analyzing the coherence of hidden state activations across altered model checkpoints. This method aligns directly with our brain score framework by operating on the same layer activations used for neural alignment analysis. The hypothesis underlying this method is that text from an individual with AD will produce more coherent activation patterns in an appropriately altered model than in either the baseline or more severely altered models, creating an "internal resonance" between input characteristics and model state.

3.3.5 Personalized AD monitoring

To evaluate the potential of altered LMs as DTs, we will implement a preliminary study using longitudinal data — ADReSSo (Luz et al., 2021) and Pitt Corpus (Becker et al., 1994) are good fits. The dataset will provide speech recordings and cognitive assessments from subjects tracked over multiple time points, enabling us to model cognitive changes longitudinally. Our approach includes three key steps: (I) selecting the initial reference model that best matches each individual's baseline cognitive state using our detection pipeline; (II) performing personalized fine-tuning with the individual's transcripts to create an individual-specific model; and (III) tracking temporal progression by updating and comparing the personalized model against reference models at each time point to map the trajectory of cognitive decline. This personalized approach allows us to create cognitive DTs that capture individual-specific language patterns rather than only population-level features, potentially enabling more sensitive detection of subtle changes in cognitive function before they become apparent through traditional assessment methods.

4 Conclusion

The solution offers two key advantages: (I) A non-invasive detection framework that could be integrated into existing clinical workflows or remote screening tools. (II) An interpretable model architecture whose behavior approximates the cognitive degradation associated with AD. This alignment could enable researchers to simulate brain-like lan-

guage processing and use the model as a reference for low-cost, virtual experiments.

Limitations

This work will have several limitations that will frame the interpretation of its results. We will not use task-matched fMRI from AD patients; normative brain models will be trained on healthy listeners. Consequently, brain-score deltas will quantify deviation from a healthy reference rather than disease-specific effects, and generalization across cohorts, languages, and tasks will remain uncertain. Additionally, inter-individual variability in fMRI responses presents a challenge; we will leverage datasets providing both spatially smoothed and non-smoothed outputs (Nastase et al., 2021), as smoothing can improve cross-subject alignment at the expense of spatial specificity. Moreover, brain-model alignment will be correlational: brain score will rely on linear mappings, so high predictivity will not imply mechanistic equivalence, and layer-ROI associations will be interpreted cautiously. The MoCA-inspired protocol will be adapted to text-only interaction and will omit visiospatial items; scores will therefore reflect task performance under this interface rather than a clinical diagnosis and may depend on model size, prompting, and calibration.

The alteration strategies will approximate aspects of language decline and cognitive impairment, but their capacity to act as faithful disease models remains to be established. If alterations do not adequately reflect the neurobiological changes associated with AD, or if brain alignment metrics prove insufficiently sensitive to capture cognitive decline nuances, this would reveal fundamental constraints in current approaches. Similarly, while alterations may measurably impact cognitive performance, they may not necessarily enhance classification performance, potentially indicating that effective detection requires balancing simulation realism with practical utility. For personalized monitoring, the amount of available individual data may be insufficient to capture meaningful longitudinal changes, and inter-individual variability in language patterns may limit the effectiveness of few-shot adaptation for progression tracking.

Nevertheless, even if primary hypotheses are not fully supported, the study will yield valuable insights. The comparative nature of our framework ensures that differential patterns—such as linguistic metrics showing stronger correlations with brain alignment than cognitive tasks, or certain alteration methods proving more effective than others—would provide actionable guidance for future research. The analysis across multiple alteration approaches and assessment dimensions will identify which methods are more biologically plausible and which evaluation strategies are most sensitive to cognitive changes. Even uniformly weak relationships would be informative: they would suggest that the neural alignment observed in prior studies for healthy language processing does not extend straightforwardly to pathological conditions, revealing fundamental limitations in current LM architectures for modeling disease states. Such findings would establish boundaries of LM capabilities in simulating human cognitive processes, guide future research toward more biologically plausible methods, inform which evaluation strategies are most appropriate for cognitive decline assessment, and help design next-generation models that balance simulation fidelity with practical detection performance.

Acknowledgments

The author(s) would like to thank the anonymous colleagues and mentors whose feedback and discussions contributed to improving this proposal.

References

Maitha Alarjani and Badar Almarri. 2024. fMRI-based Alzheimer's disease detection via functional connectivity analysis: A systematic review. *PeerJ Computer Science*, 10:e2302.

Ingrid Arevalo-Rodriguez, Nadja Smailagic, Marta Roqué-Figuls, Agustín Ciapponi, Erick Sanchez-Perez, Antri Giannakou, Olga L Pedraza, Xavier Bonfill Cosp, and Sarah Cullum. 2021. Mini-Mental State Examination (MMSE) for the early detection of dementia in people with mild cognitive impairment (MCI). *The Cochrane Database of Systematic Reviews*, 2021(7):CD010783.

Taniya Ashraf, Mohammad Ahsan Chisti, and Mohamed Mahees Raheem. 2024. Digital Twin for Neurology: An Introduction to a New Frontier in Healthcare. In *2024 21st Learning and Technology Conference (L&T)*, pages 284–289.

Aparna Balagopalan, Benjamin Eyre, Jessica Robin, Frank Rudzicz, and Jekaterina Novikova. 2021. Comparing Pre-trained and Feature-Based Models for Prediction of Alzheimer's Disease Based on Speech. *Frontiers in Aging Neuroscience*, 13.

Aparna Balagopalan, Benjamin Eyre, Frank Rudzicz, and Jekaterina Novikova. 2020. To bert or not to bert: Comparing speech and language-based approaches for alzheimer's disease detection. In *Interspeech 2020*, pages 2167–2171.

James T. Becker, François Boiler, Oscar L. Lopez, Judith Saxton, and Karen L. McGonigle. 1994. The natural history of alzheimer's disease: Description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594.

Marcel Binz and Eric Schulz. Using cognitive psychology to understand GPT-3. 120(6):e2218523120.

Charlotte Caucheteux. 2023. *Language Representations in Deep Learning Algorithms and the Brain*. Theses, Université Paris-Saclay.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Remi King. 2021. Disentangling syntax and semantics in the brain with deep networks. In *Proceedings of the 38th International Conference on Machine Learning*, pages 1336–1348. PMLR.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. 2022. Deep language algorithms predict semantic comprehension from brain activity. *Scientific Reports*, 12(1):16327.

Charlotte Caucheteux and Jean-Rémi King. 2021. Language processing in brains and deep neural networks: Computational convergence and its limits.

Jungho Cha, Hang Joon Jo, Hee Jin Kim, Sang Won Seo, Han-Soo Kim, Uicheul Yoon, Hyunjin Park, Duk L. Na, and Jong-Min Lee. 2013. Functional alteration patterns of default mode networks: Comparisons of normal aging, amnestic mild cognitive impairment and Alzheimer's disease. *The European Journal of Neuroscience*, 37(12):1916–1924.

Lei Chi, Arav Sharma, Ari Gebhardt, and Joseph T. Colonel. 2025. Predicting Cognitive Decline: A Multimodal AI Approach to Dementia Screening from Speech. *Preprint*, arXiv:2502.08862.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Roy Dayan, Benjamin Uliel, and Gal Koplewitz. 2024. Age against the machine—susceptibility of large language models to cognitive impairment: Cross sectional analysis. *BMJ*, 387:e081948.

Emily L. Dennis and Paul M. Thompson. 2014. Functional Brain Connectivity using fMRI in Aging and Alzheimer's Disease. *Neuropsychology review*, 24(1):49–62.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Preprint*, arXiv:1810.04805.

Kewen Ding, Madhu Chetty, Azadeh Noori Hoshyar, Tanusri Bhattacharya, and Britt Klein. 2024. Speech based detection of Alzheimer's disease: A survey of AI techniques, datasets and challenges. *Artificial Intelligence Review*, 57(12):325.

Kimiko Domoto-Reilly, Daisy Sapolsky, Michael Brickhouse, and Bradford C. Dickerson. 2012. Naming impairment in Alzheimer's disease is associated with left anterior temporal lobe atrophy. *NeuroImage*, 63(1):348–355.

Kathleen C. Fraser, Jed A. Meltzer, and Frank Rudzicz. 2016. Linguistic Features Identify Alzheimer's Disease in Narrative Speech. *Journal of Alzheimer's disease: JAD*, 49(2):407–422.

Giovanni B. Frisoni, Jean-Marie Annoni, Stefanie Becker, Tim Brockmann, Markus Buerge, Jean-François Démonet, Dan Georgescu, Anton Gietl, Ulrich Hemmeter, Stefan Klöppel, Thomas Leyhe, Andreas U. Monsch, Franco Rogantini, Delphine Roulet Schwab, Egemen Savaskan, Karl Schaller, Armin von Gunten, and Gabriel Gold. 2021. Position Statement on Anti-Dementia Medication for Alzheimer's Disease by Swiss Stakeholders. *Clinical and Translational Neuroscience*, 5(2):14.

Jon Gauthier and Roger Levy. 2019. Linking artificial and human neural representations of language. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 529–539, Hong Kong, China. Association for Computational Linguistics.

Joshua T. Goodman. 2001. A bit of progress in language modeling. *Computer Speech & Language*, 15(4):403–434.

Michael D. Greicius, Gaurav Srivastava, Allan L. Reiss, and Vinod Menon. 2004. Default-mode network activity distinguishes Alzheimer's disease from healthy aging: Evidence from functional MRI. *Proceedings of the National Academy of Sciences of the United States of America*, 101(13):4637–4642.

Murray Grossman, Phyllis Koenig, Guila Glosser, Chris DeVita, Peachie Moore, Jina Rhee, John Detre, David Alsop, Jim Gee, and fMRI study. Functional magnetic resonance imaging. 2003. Neural basis for semantic memory difficulty in Alzheimer's disease: An fMRI study. *Brain: A Journal of Neurology*, 126(Pt 2):292–311.

Laura Hernández-Domínguez, Sylvie Ratté, Gerardo Sierra-Martínez, and Andrés Roche-Bergua. 2018. Computer-based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task. *Alzheimer's & Dementia (Amsterdam, Netherlands)*, 10:260–268.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9:1735–1780.

Alexander G. Huth, Tyler Lee, Shinji Nishimoto, Natalia Y. Bilenko, An T. Vu, and Jack L. Gallant. 2016. Decoding the Semantic Content of Natural Movies from Human Brain Activity. *Frontiers in Systems Neuroscience*, 10:81.

Aneta Kielar, Tiffany Deschamps, Ron K. O. Chu, Regina Jokel, Yasha B. Khatamian, Jean J. Chen, and Jed A. Meltzer. 2016. Identifying Dysfunctional Cortex: Dissociable Effects of Stroke and Aging on Resting State Dynamics in MEG and fMRI. *Frontiers in Aging Neuroscience*, 8.

Alexandra König, Aharon Satt, Alexander Sorin, Ron Hoory, Orith Toledo-Ronen, Alexandre Derreumaux, Valeria Manera, Frans Verhey, Pauline Aalten, Phillipe H. Robert, and Renaud David. 2015. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimer's & Dementia (Amsterdam, Netherlands)*, 1(1):112–124.

Alyssa M. Lanzi, Anna K. Saylor, Davida Fromm, Houjun Liu, Brian MacWhinney, and Matthew L. Cohen. 2023. DementiaBank: Theoretical Rationale, Protocol, and Illustrative Analyses. *American Journal of Speech-Language Pathology*, 32(2):426–438. Publisher: American Speech-Language-Hearing Association.

R. Laubenbacher, B. Mehrad, I. Shmulevich, and N. Trayanova. 2024. Digial Twins in Medicine. *Nature computational science*, 4(3):184–191.

Changye Li, David Knopman, Weizhe Xu, Trevor Cohen, and Serguei Pakhomov. 2022a. GPT-D: Inducing Dementia-related Linguistic Anomalies by Deliberate Degradation of Artificial Neural Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1866–1877, Dublin, Ireland. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Jixing Li, Shohini Bhattasali, Shulin Zhang, Berta Franzluebbers, Wen-Ming Luh, R. Nathan Spreng, Jonathan R. Brennan, Yiming Yang, Christophe Pallier, and John Hale. 2022b. Le Petit Prince multilingual naturalistic fMRI corpus. *Scientific Data*, 9(1):530.

Xiaohui Liang, John A. Batsis, Youxiang Zhu, Tiffany M. Driesse, Robert M. Roth, David Kotz, and Brian MacWhinney. 2022. Evaluating voice-assistant commands for dementia detection. *Computer Speech & Language*, 72:101297.

Nikos K. Logothetis, Jon Pauls, Mark Augath, Torsten Trinath, and Axel Oeltermann. 2001. Neurophysiological investigation of the basis of the fMRI signal. *Nature*, 412(6843):150–157.

Saturnino Luz, Sofia De La Fuente Garcia, Fasih Haider, Davida Fromm, Brian MacWhinney, Alyssa Lanzi, Ya-Ning Chang, Chia-Ju Chou, and Yi-Chien Liu. 2024. Connected Speech-Based Cognitive Assessment in Chinese and English. pages 947–951.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020. Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge. *Preprint*, arXiv:2004.06833.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2021. Detecting cognitive decline using speech only: The ADReSSo Challenge. *Preprint*, arXiv:2104.09356.

Alan M. Mandell and Robert C. Green. 2011. Alzheimer's Disease. In *The Handbook of Alzheimer's Disease and Other Dementias*, chapter 1, pages 1–91. John Wiley & Sons, Ltd.

Mohammad Momenian, Zhengwu Ma, Shuyi Wu, Chengcheng Wang, Jonathan Brennan, John Hale, Lars Meyer, and Jixing Li. 2024. Le Petit Prince Hong Kong (LPPHK): Naturalistic fMRI and EEG data from older Cantonese speakers. *Scientific Data*, 11(1):992.

Ziad S. Nasreddine, Natalie A. Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L. Cummings, and Howard Chertkow. 2005. The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool For Mild Cognitive Impairment. *Journal of the American Geriatrics Society*, 53(4):695–699.

Samuel A. Nastase, Yun-Fei Liu, Hanna Hillman, Asieh Zadbood, Liat Hasenfratz, Neggin Keshavarzian, Janice Chen, Christopher J. Honey, Yaara Yeshurun, Mor Regev, Mai Nguyen, Claire H. C. Chang, Christopher Baldassano, Olga Lositsky, Erez Simony, Michael A. Chow, Yuan Chang Leong, Paula P. Brooks, Emily Micciche, and 6 others. 2021. The "Narratives" fMRI dataset for evaluating models of naturalistic language comprehension. *Scientific Data*, 8(1):250.

Swati Padhee, Anurag Illendula, Megan Sadler, Valerie L. Shalin, Tanvi Banerjee, Krishnaprasad Thirunarayan, and William L. Romine. 2020. Predicting early indicators of cognitive decline from verbal utterances. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 477–480.

Yilin Pan, Bahman Mirheidari, Jennifer M. Harris, Jennifer C. Thompson, Matthew Jones, Julie S. Snowden, Daniel Blackburn, and Heidi Christensen. 2021.

Using the Outputs of Different Automatic Speech Recognition Paradigms for Acoustic- and BERT-Based Alzheimer's Dementia Detection Through Spontaneous Speech. In *Proc. Interspeech 2021*, pages 3810–3814.

Silvia Pelucchi, Fabrizio Gardoni, Monica Di Luca, and Elena Marcello. 2022. Synaptic dysfunction in early phases of Alzheimer's Disease. *Handbook of Clinical Neurology*, 184:417–438.

Xiaoke Qi, Qing Zhou, Jian Dong, and Wei Bao. 2023. Noninvasive automatic detection of Alzheimer's disease from spontaneous speech: A review. *Frontiers in Aging Neuroscience*, 15:1224723.

Alec Radford, Jeff Wu, R. Child, D. Luan, Dario Amodei, and I. Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Matthew A. Lambon Ralph, Elizabeth Jefferies, Karalyn Patterson, and Timothy T. Rogers. 2017. The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, 18(1):42–55.

Saige Rutherford, Seyed Mostafa Kia, Thomas Wolfers, Charlotte Fraza, Mariam Zabihi, Richard Dinga, Pierre Berthet, Amanda Worker, Serena Verdi, Henricus G. Ruhe, Christian F. Beckmann, and Andre F. Marquand. 2022. The normative modeling framework for computational psychiatry. *Nature Protocols*, 17(7):1711–1734.

Martin Schrimpf, Idan Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua Tenenbaum, and Evelina Fedorenko. 2020. Artificial Neural Networks Accurately Predict Language Processing in the Brain.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. 2018. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv : the preprint server for biology*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Robert D. Terry, Eliezer Masliah, David P. Salmon, Nelson Butters, Richard DeTeresa, Robert Hill, Lawrence A. Hansen, and Robert Katzman. 1991. Physical basis of cognitive alterations in alzheimer's disease: Synapse loss is the major correlate of cognitive impairment. *Annals of Neurology*, 30(4):572–580.

C. Thomas, V. Keselj, N. Cercone, K. Rockwood, and E. Asp. 2005. Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. In *IEEE International Conference Mechatronics and Automation, 2005*, volume 3, pages 1569–1574 Vol. 3.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need. *Preprint*, arXiv:1706.03762.

M. Verma and R. J. Howard. 2012. Semantic memory and language dysfunction in early Alzheimer's disease: A review. *International Journal of Geriatric Psychiatry*, 27(12):1209–1217.

Eric Williams, Megan McAuliffe, and Catherine Theys. 2021. Language changes in Alzheimer's disease: A systematic review of verb processing. *Brain and Language*, 223:105041.

Adrian Wong, David Nyenhuis, Sandra E. Black, Lorraine S. N. Law, Eugene S. K. Lo, Pauline W. L. Kwan, Lisa Au, Anne Y. Y. Chan, Lawrence K. S. Wong, Ziad Nasreddine, and Vincent Mok. 2015. Montreal Cognitive Assessment 5-minute protocol is a brief, valid, reliable, and feasible cognitive screen for telephone administration. 46(4):1059–1064.

Jiqing Wu and Viktor H. Koelzer. 2024. Towards generative digital twins in biomedical research. *Computational and Structural Biotechnology Journal*, 23:3481–3488.

Qin Yang, Xin Li, Xinyun Ding, Feiyang Xu, and Zhenhua Ling. 2022. Deep learning-based speech analysis for Alzheimer's disease detection: A literature review. *Alzheimer's Research & Therapy*, 14(1):186.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.